600309

BOOZ • ALLEN APPLIED RESEARCH INC.

MAY [illegible]7 1964

Best Available Copy

Bivariate Regression When Both Variables Are Random

Informal Report No. 3

to

Systems Analysis Division
Chemical Research and Development Laboratories
Edgewood Arsenal, Maryland

under

Contract DA-18-108-CML-7174

May 1964 803-1-R15

BOOZ·ALLEN APPLIED RESEARCH, INC.

CHICAGO
WASHINGTON

BIVARIATE REGRESSION WHEN BOTH VARIABLES ARE RANDOM

In dealing with certain estimation problems in biological and chemical research it is frequently necessary to compute a regression equation which can be used to predict values of a variable Y for selected values of another variable X. The standard procedure calls for selecting a fixed set of values of X and then sampling Y. If this procedure is followed, then the resulting regression equation is

$$Y' = a_1 + b_1 X, \tag{1}$$

where

$$b_1 = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \Big/ \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{2}$$

$$a_1 = \overline{Y} - b_1 \overline{X}. \tag{3}$$

This family of equations is classically used in computing the regression equation for Y on X.

If, on the other hand, it is desired to select a set of Y values, sample X and then construct the regression function for X on Y, the resulting equation is

$$X' = a_2 + b_2 Y, \tag{4}$$

where

$$b_2 = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \Big/ \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{5}$$

$$a_2 = \overline{X} - b_2 \overline{Y}. \tag{6}$$

The above formulas are obtained using the method of least squares. Furthermore, if one is willing to assume that for the first situation Y is normally distributed with a common variance about the regression line, and for the second situation X is normally distributed with a common variance about the regression line, then the estimates above are also maximum likelihood estimates.

Unfortunately, in practice it is impossible always to control the independent variable X or Y, as the case may be. In these situations, then, both variables will be subject to error, or random variation. (For example, in estimating a dose-response function, both the dose and the proportion responding to that dose may be random variables, since dose frequently cannot be measured precisely.) When such a sampling situation arises it seems advisable to consider using orthogonal regression

techniques. In this case the sum of squares of the perpendicular distances to the regression line will be minimized. If, furthermore, the variances for X and Y can be standardized in some sense, then the estimates which follow are also maximum likelihood estimates. The following equations assume that X will be used to predict Y. An interchange of the X and Y values will make it possible to arrive at the equation for predicting X from Y. The necessary formulas are:

$$Y' = a_3 + b_3 X, \tag{7}$$

where

$$b_3 = \frac{\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i^2 + \left\{ \left(\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i^2 \right)^2 + 4 \left(\sum_{i=1}^{n} x_i y_i \right)^2 \right\}^{1/2}}{2 \sum_{i=1}^{n} x_i y_i} \tag{8}$$

$$a_3 = \overline{Y} - b_3 \overline{X}, \tag{9}$$

and

$$y_i = (Y_i - \overline{Y}),\ x_i = (X_i - \overline{X}). \tag{10}$$

(Two references on estimation when both variables are subject to random error appear at the end of this memo.)

An Example

To illustrate the kinds of results which can be obtained, the following example has been selected from Reference 3. The data represent the heights and weights of 12 men:

X (height in inches):	60	60	60	62	62	62	62	64	64	70	70	70
Y (weight in pounds):	110	135	120	120	140	130	135	150	145	170	185	160

The four regression equations are now summarized:

(1) Regression of Y on X, standard:

$$Y' = -179.36 + 5.029\,X$$

(2) Regression of X on Y, standard:

$$X' = 40.61 + .164\,Y$$

(3) Regression of Y on X, orthogonal:

$$Y' = -245.57 + 6.066\,X$$

(4) Regression of X on Y, orthogonal:

$$X' = 40.48 + .165\,Y.$$

For this particular illustration it should be noted that the results for one case (X on Y) are quite close together, while for the other case (Y on X) the differences could be significant in any inferential treatment of the data.

Summary

This memo presents the method of orthogonal regression and compares the standard linear regression procedures with orthogonal linear regression procedures. Care should be exercised in using the standard methods when both X and Y are subject to random variation.

~~The computer program description~~ and the fortran program are attached.

REFERENCES

1. Wald, A., "The Fitting of Straight Lines if Both Variables are Subject to Error," Annals of Mathematical Statistics, Volume 11, 1940.

2. Bartlett, M.S., "Fitting a Straight Line When Both Variables are Subject to Error," Biometrics, Volume 5, 1949.

3. Dixon, W.J., and Massey, F.J., Introduction to Statistical Analysis, McGraw-Hill Book Company, Inc., Chapter 11, 1951.

A. IDENTIFICATION

Title: Orthogonal Regression

Identification:

Category:

Programmer: Freida E. Robey

Date: October, 1963

B. PURPOSE - This program computes the mean $X(\overline{X})$, mean $Y(\overline{Y})$, the correlation coefficient (R), the orthogonal regression line for Y on X and X on Y, and the sum of the minimum residuals.

C. USAGE

1. Operational Procedure

This program is in FORTRAN

(a) Machine load Compiler tape III (the interpreter) at P = 0000. Check sum - 0000.

(b) Clear, position the binary object tape in the reader and run (from P = 0000).

Error Stop: P - 0052 Parity error stop. Usually indicates punch trouble.

(c) The FORTRAN object program is in memory and ready to be executed. Turn on the punch, position input tape (data) in the reader and run (from P = 1020).

2. Data

The first value on the data tape is N: the number of pairs of data. The next values on the data tape are X_n and Y_n pairs.

Format	Definition	Example
I3	N	3/
2F20.8	X, Y	60./110./
		60./135./
		60./120./

3. Output

The output is punched in flexowriter code, and includes $\overline{X}$, $\overline{Y}$, R (the correlation coefficient), the equations of the regression lines Y on X and X on Y and the sum of the minimum residuals. The equations used for computation are:

$$\hat{b} = \frac{\left(\sum y_i^2 - \sum x_i^2\right) + \sqrt{\left[\sum y_i^2 - \sum x_i^2\right]^2 + 4\left(\sum x_i y_i\right)^2}}{2\sum x_i y_i}$$

where

$$y = Y_i - \overline{Y},$$

$$x = X_i - \overline{X},$$

$$a = \overline{Y} - \hat{b}\overline{X}.$$

Regression line Y on X is:

$$Y' = a + \hat{b}X$$

$$\hat{a} = \frac{(\sum x_i^{\,2} - \sum x_i^{\,2}) + \sqrt{\left[\sum y_i^{\,2} - \sum x_i^{\,2}\right]^2 + 4(\sum x_i y_i)^2}}{2\sum x_i y_i}$$

$$a_i = \overline{X} - \hat{a}\overline{Y}.$$

Regression line X on Y is:

$$X' = a_1 + \hat{a}Y$$

Minimum residuals:

$$\sum d_i^{\,2} = \sum_i \frac{(Y_i - a - \hat{b}X_i)^2}{1 + \hat{b}^2}$$

$$\sum d_i^{\,2} = \sum_i \frac{(X_i - a_1 - \hat{a}Y_i)^2}{1 + \hat{a}^2}.$$

```
C       ORTHOGONAL REGRESSION
10      FORMAT (13)
11      FORMAT (2F20.8)
12      FORMAT (6HXBAR=;,F14.8, 7H;YBAR=;,F14.8, 7H;;;;R=;,F14.8/)
13      FORMAT (20HREGRESSION; EQUATIONS/)
14      FORMAT (17HREGRESSION; Y;ON;X/)
15      FORMAT (8HYPRIME=;,F14.8, 4H;+;;,F14.8,1HX/)
16      FORMAT (17HREGRESSION;X;ON;Y/)
17      FORMAT (8HXPRIME=;,F14.8,4H;+;;,F14.8,1HY/)
18      FORMAT (17MINIMUM;RESIDUALS/)
19      FORMAT (19HSUM;D(I);SQUARED;=;,F14.8/)
        DIMENSION X(100), Y(100)
1       XES=0
        YES=0
        SUMX=0
        SUMY=0
        SUMXY=0
        SUMIN1=0
        SUMIN2=0
        READ 10,N
        READ 11,(X(I), Y(I),I=1,N)
C       COMPUTE SUMS
        DO 20 I=1,N
        XES=XES+X(I)
20      YES=YES+Y(I)
        XBAR=XES/N
        YBAR=YES/N
        DO 25 I=1,N
        Y(I)=Y(I)-YBAR
        X(I)=X(I)-XBAR
        SUMX=SUMX+X(I)'X(I)
        SUMY=SUMY+Y(I)'Y(I)
25      SUMXY=SUMXY+X(I)'Y(I)
C       COMPUTE REGRESSION COEFFS
        DIFSSQ=SUMY-SUMX
        RADCAL=SQRTF(DIFSSQ'DIFSSQ+4.'SUMXY'SUMXY)
        DENOM=2.'SUMXY
        BHAT=(DIFSSQ+RADCAL)/DENOM
        A=YBAR-BHAT'XBAR
        DIFSSQ=SUMX-SUMY
        AHAT=(DIFSSQ+RADCAL)/DENOM
        A1=XBAR-AHAT'YBAR
C       COMPUTE R
        R=SQRTF(BHAT'AHAT)
```

```
C      COMPUTE MINIMUM RESIDUALS
       DENOM=1.+BHAT'BHAT
       DO 40 I=1,N
       Y(I)=Y(I)+YBAR
       X(I)=X(I)+XBAR
       BNUM=Y(I)-A-BHAT'X(I)
       BNUM2=BNUM1+BNUM2
40     SUMIN1=SUMIN1+BNUM2
       SUMIN1=SUMIN1/DENOM
       DENOM1=1.+AHAT'AHAT
       DO 50 I=1,N
       ANUM=X(I)-A1_AHAT'Y(I)
       ANUM2=ANUM'ANUM
50     SUMIN2=SUMIN2+ANUM2
       SUMIN2=SUMIN2/DENOM1
       PUNCH 12,XBAR, YBAR, R
       PUNCH 13
       PUNCH 14
       PUNCH 15, A, BHAT
       PUNCH 16
       PUNCH 17, A1, AHAT
       PUNCH 18
       PUNCH 19, SUMIN1
       PUNCH 19, SUMIN2
       PAUSE 0001
       GO TO 1
       END
       END
```